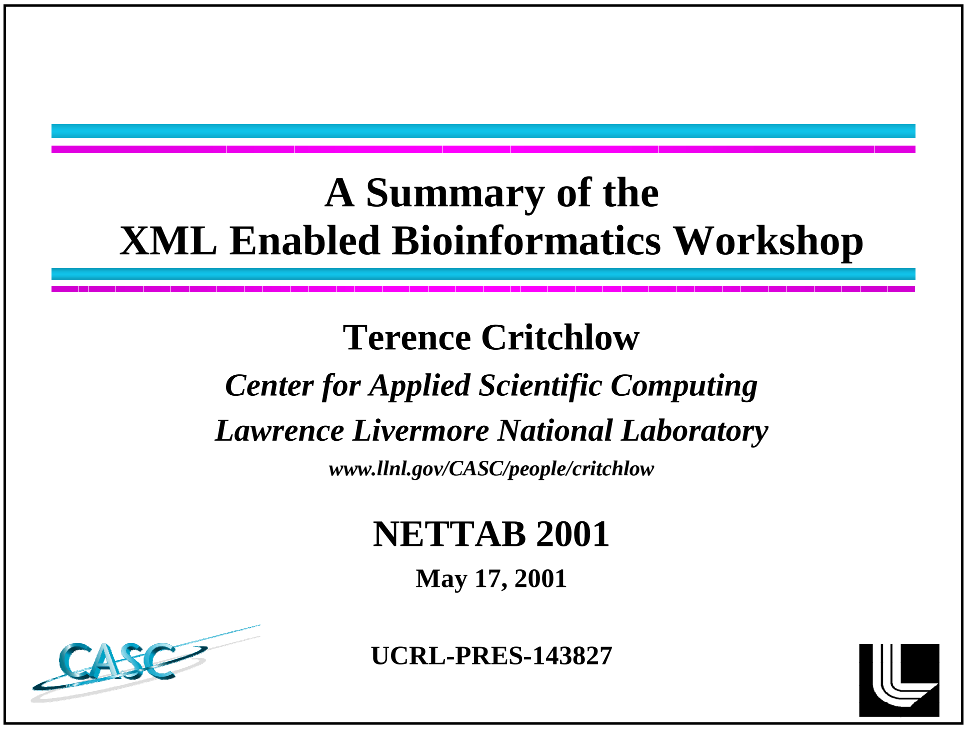


A Summary of the XML Enabled Bioinformatics Workshop

Terence Critchlow

Center for Applied Scientific Computing

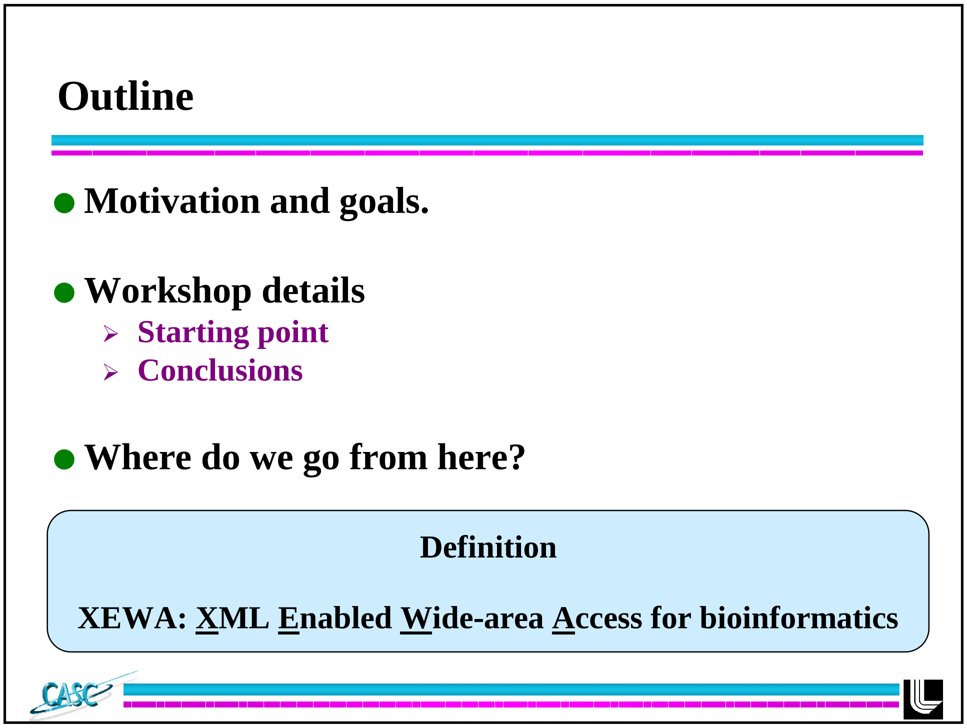
Lawrence Livermore National Laboratory

www.llnl.gov/CASC/people/critchlow

NETTAB 2001

May 17, 2001

UCRL-PRES-143827



Outline

- **Motivation and goals.**
- **Workshop details**
  - **Starting point**
  - **Conclusions**
- **Where do we go from here?**

Definition

XEWA: XML Enabled Wide-area Access for bioinformatics

What is XEWA about?

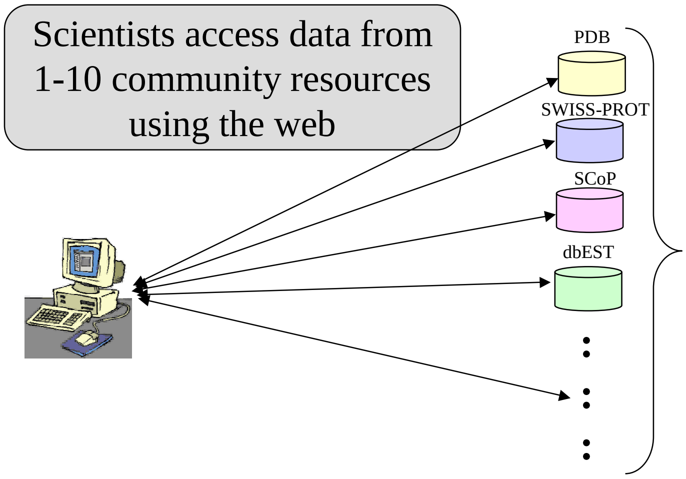
- XEWA started as a two-day “working” workshop sponsored by IEEE.
  - XEWA is now an ongoing initiative at LLNL
  - The workshop focus was not on presentations, but rather on breakout sessions which attempted to address aspects of a specific problem facing bioinformatics.

Motivation

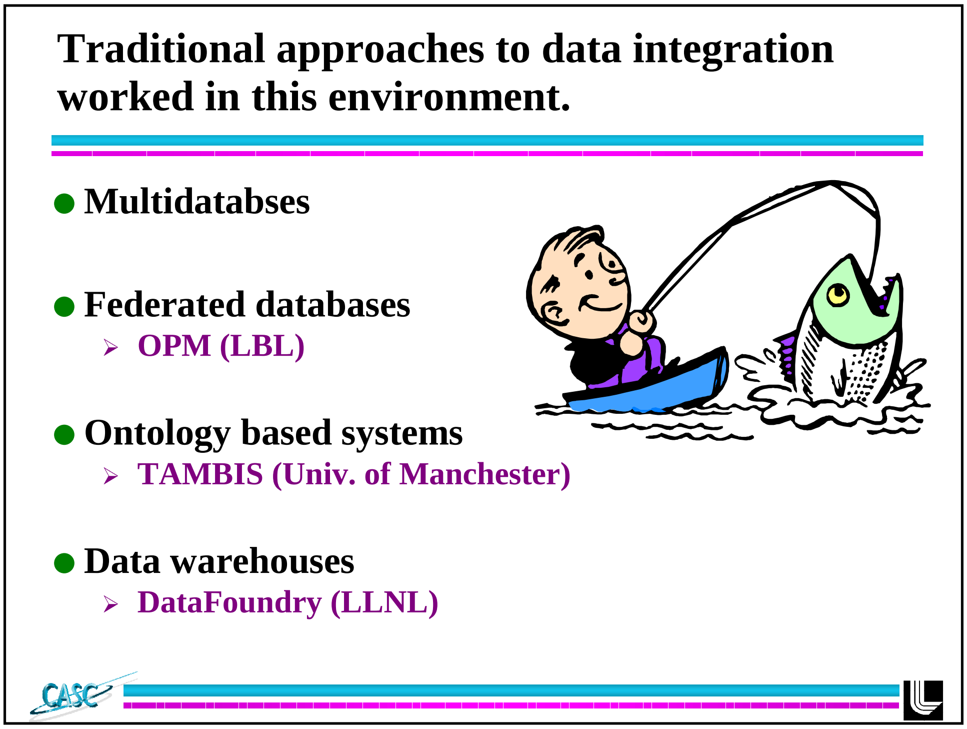
The gap between the information available to scientists and the information they actually use is growing.

Four years ago, there were a limited number of bioinformatics data sources.

Scientists access data from 1-10 community resources using the web

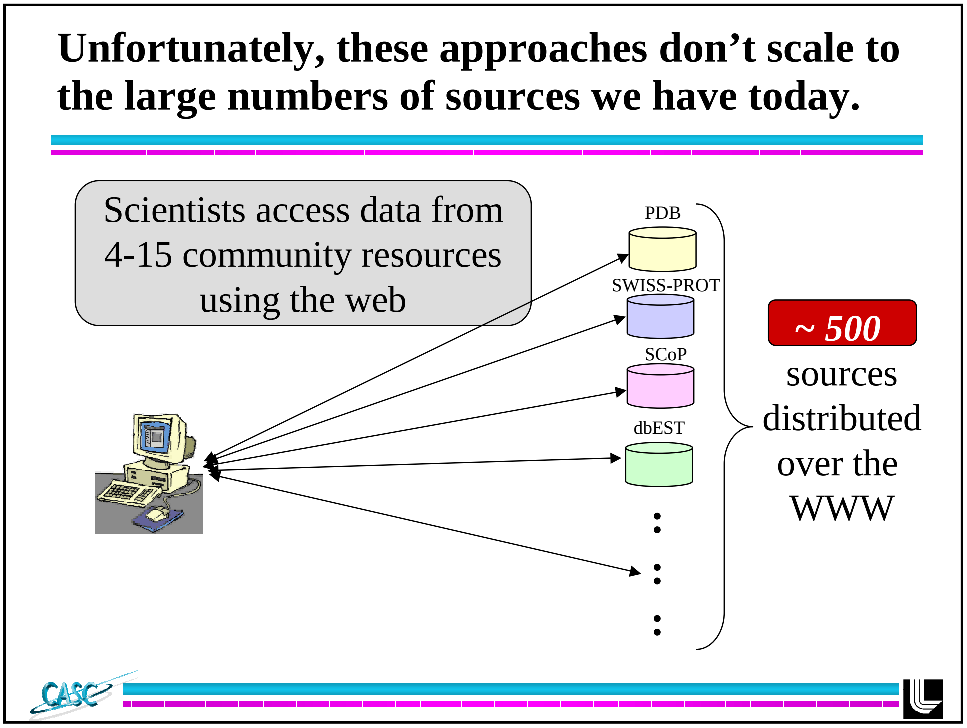


~ 30 sources distributed over the WWW



Traditional approaches to data integration worked in this environment.

- Multidatabses
- Federated databases
  - OPM (LBL)
- Ontology based systems
  - TAMBIS (Univ. of Manchester)
- Data warehouses
  - DataFoundry (LLNL)



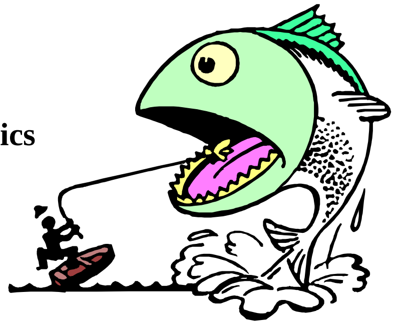
Unfortunately, these approaches don't scale to the large numbers of sources we have today.

Scientists access data from 4-15 community resources using the web

~ 500 sources distributed over the WWW

So, what makes this so much more difficult?

- No central directory
  - How do you find the sources in the first place?
- Different data formats and semantics
  - Once you find a source, how to you make sense of it?
- Complex analysis
  - Queries require more than simple data retrieval, they need to invoke complex programs.



- The problem does not scale
  - Keep up with changing interfaces and data formats
  - Need to select appropriate subset for each user / query.
  - How do you present the results in an useful format?
  - Too much work to be done manually.

XEWA is a first step at addressing the issues in large-scale bioinformatics data integration.

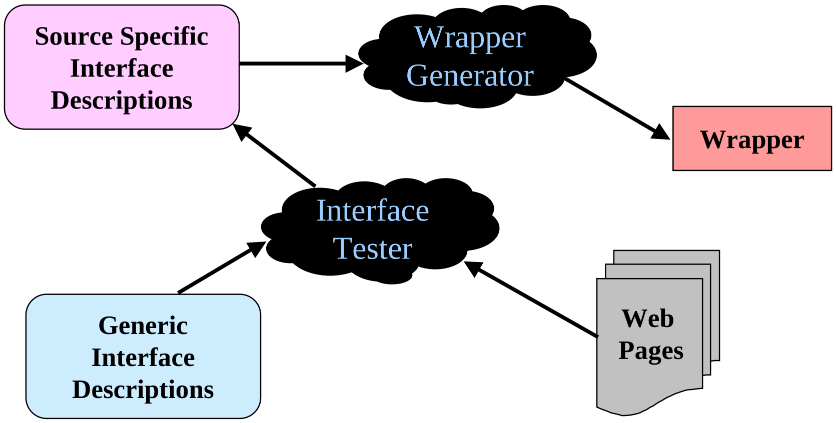
Goal

Develop an XML-based representation capable of describing how to interact with bioinformatics data sources.

Why?

Having a description of how to interact with sources simplifies building tools on top of them.

Wrappers could be automatically generated from interface descriptions.

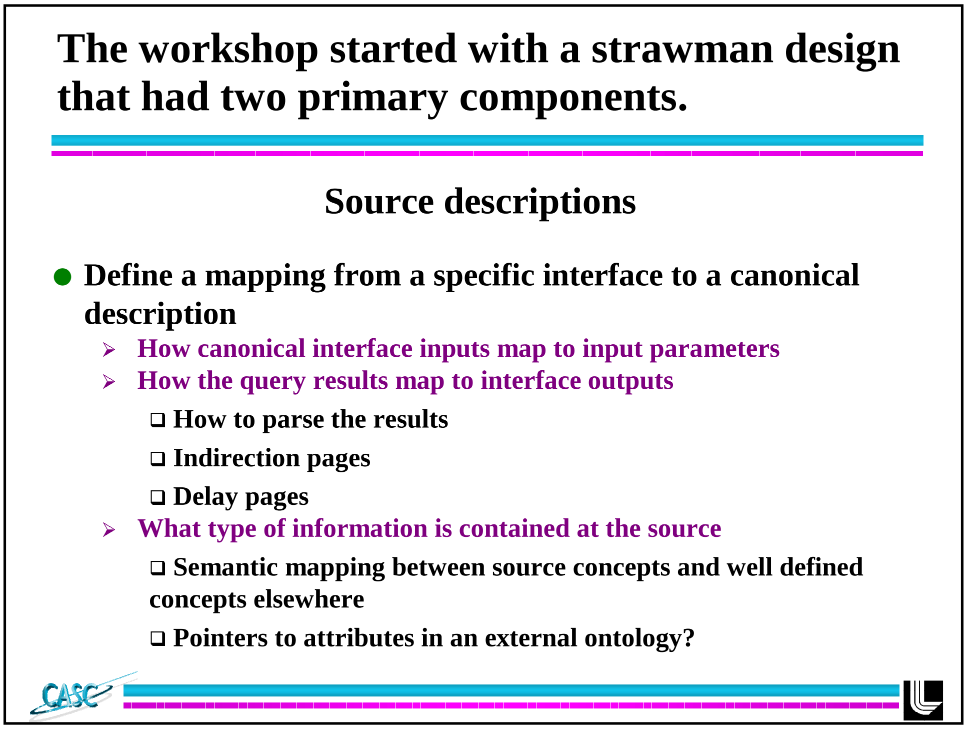

Source Specific Interface Descriptions
Wrapper Generator
Wrapper
Interface Tester
Generic Interface Descriptions
Web Pages

The workshop started with a strawman design that had two primary components.

Service classes

- Attributes
  - Represent common characteristics in the domain
  - Datatype may include a regular expression
  - Examples
- Canonical interfaces
  - Groups attributes into typical interfaces
  - Optional and required attributes
- Constraints
  - A protein sequence is not a DNA sequence
  - If a database is a DNA database then program can't be blastp

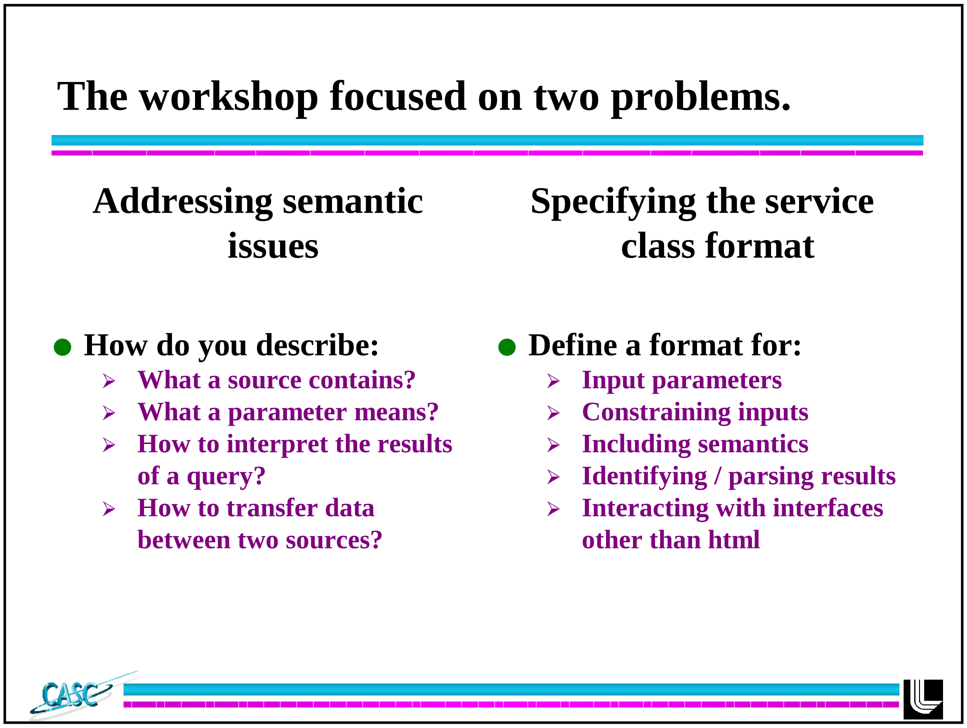
Related to both attributes and interfaces.



The workshop started with a strawman design that had two primary components.

Source descriptions

- Define a mapping from a specific interface to a canonical description
  - How canonical interface inputs map to input parameters
  - How the query results map to interface outputs
    - ❑ How to parse the results
    - ❑ Indirection pages
    - ❑ Delay pages
  - What type of information is contained at the source
    - ❑ Semantic mapping between source concepts and well defined concepts elsewhere
    - ❑ Pointers to attributes in an external ontology?



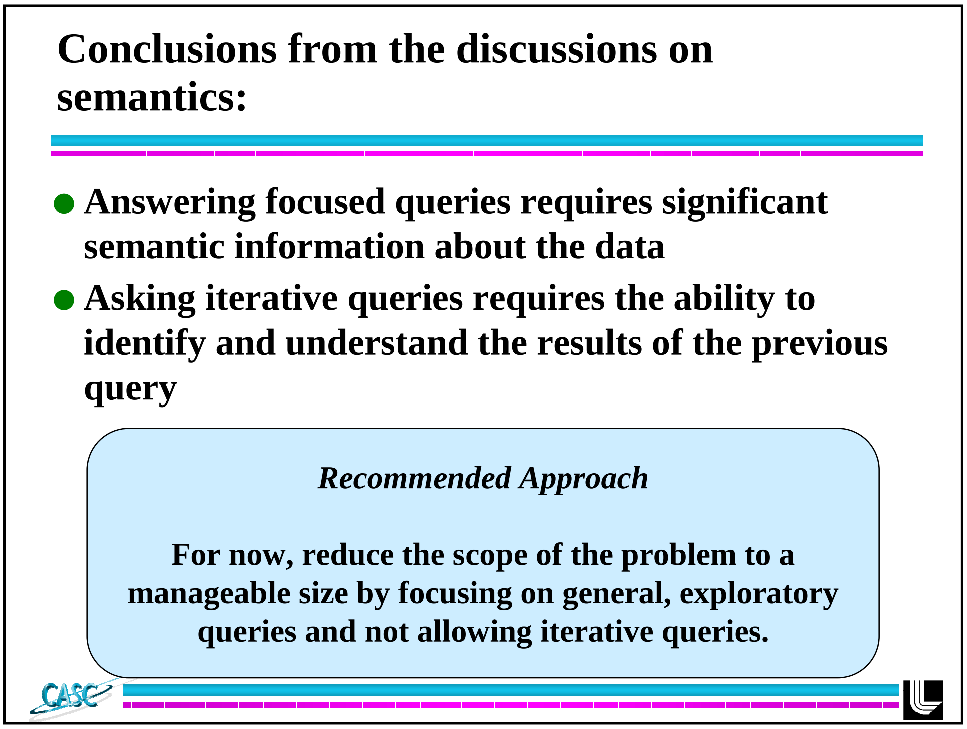
The workshop focused on two problems.

Addressing semantic issues

- How do you describe:
  - What a source contains?
  - What a parameter means?
  - How to interpret the results of a query?
  - How to transfer data between two sources?

Specifying the service class format

- Define a format for:
  - Input parameters
  - Constraining inputs
  - Including semantics
  - Identifying / parsing results
  - Interacting with interfaces other than html

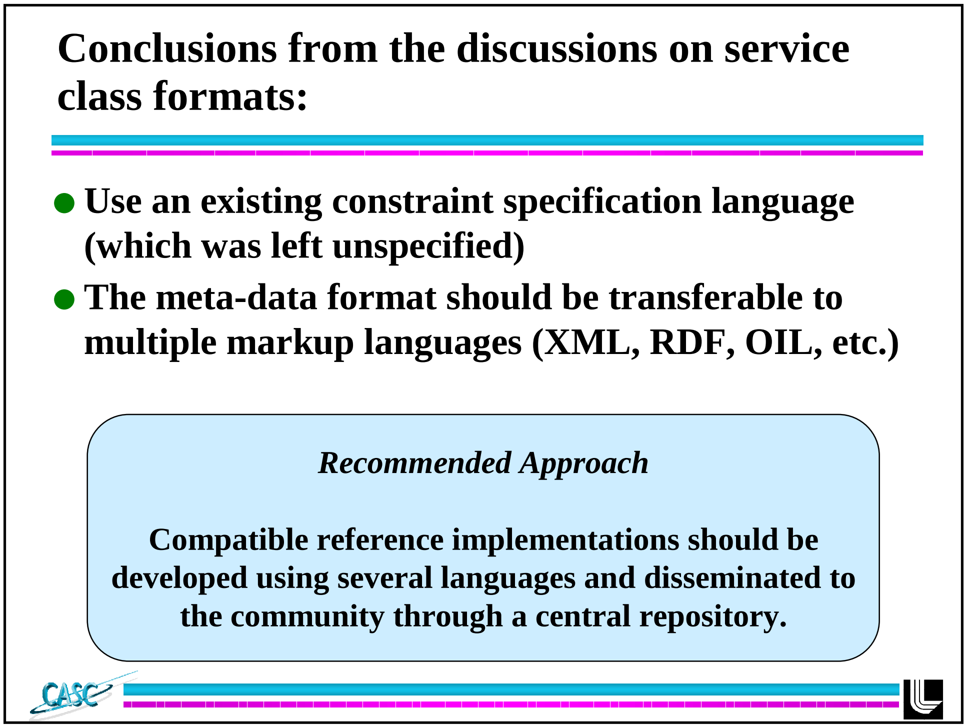


Conclusions from the discussions on semantics:

- Answering focused queries requires significant semantic information about the data
- Asking iterative queries requires the ability to identify and understand the results of the previous query

Recommended Approach

For now, reduce the scope of the problem to a manageable size by focusing on general, exploratory queries and not allowing iterative queries.



Conclusions from the discussions on service class formats:

- Use an existing constraint specification language (which was left unspecified)
- The meta-data format should be transferable to multiple markup languages (XML, RDF, OIL, etc.)

Recommended Approach

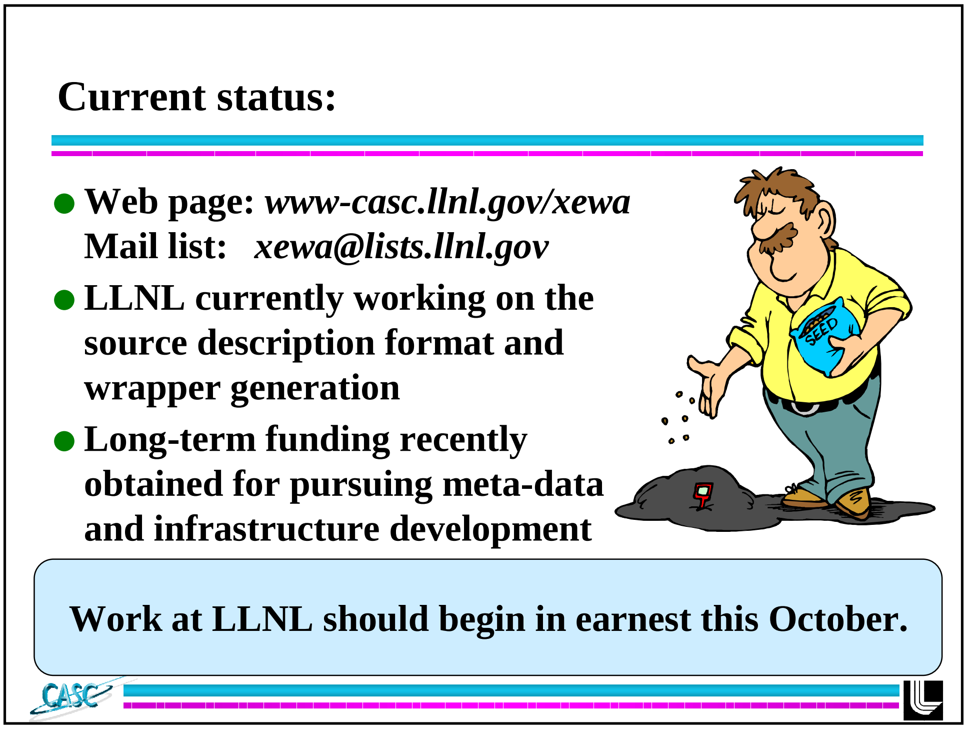
Compatible reference implementations should be developed using several languages and disseminated to the community through a central repository.

An example service class description:

```
<service_class>  <name> Blast Service </name> <input_parameters>
<attribute> <required /> <name> sequence </name>
   <type>  <union>
          <type> <class> GenericSequence</class> </type>
          <type><class> Annotationnumber</class> </type>
          <type><class> FileName</class></type></union>
   </type></attribute>
<attribute> <required/> <name> outputFormat </name>
   <type> <enum><item> email </item> <item> http </item></enum>
   </type></attribute>
<attribute> <optional/> <name> filter </name>
   <type> <boolean /> </type> </attribute>
…</output_parameters></service>
```

Where do we go from here?

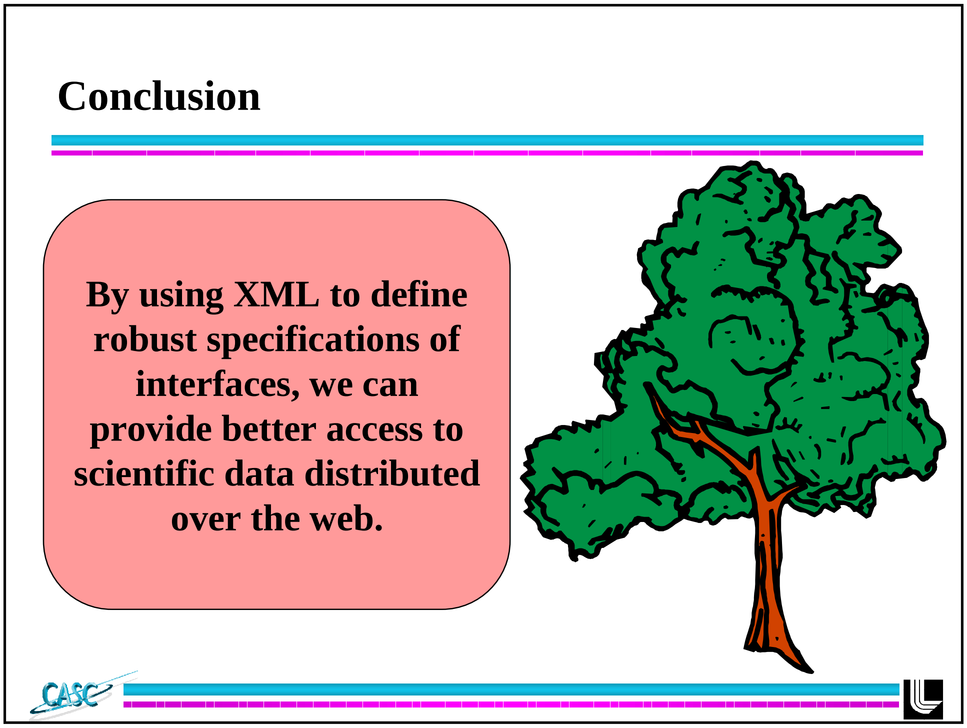
- Action items:
  - Establish forum for follow-up discussions & information exchange
  - Develop service class reference models
  - Create repository for service class descriptions
  - Implement small-scale demo for ISMB-2002
  - Hold second workshop co-located with ISMB-2002



Current status:

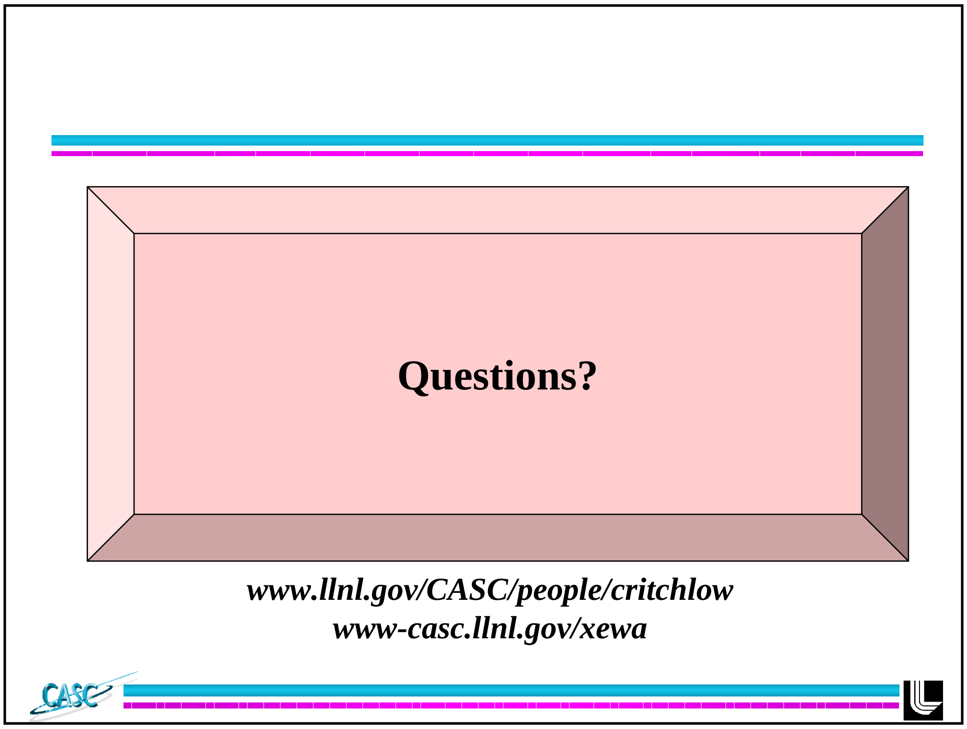
- Web page: *www-casc.llnl.gov/xewa*
  Mail list: *xewa@lists.llnl.gov*
- LLNL currently working on the source description format and wrapper generation
- Long-term funding recently obtained for pursuing meta-data and infrastructure development

Work at LLNL should begin in earnest this October.



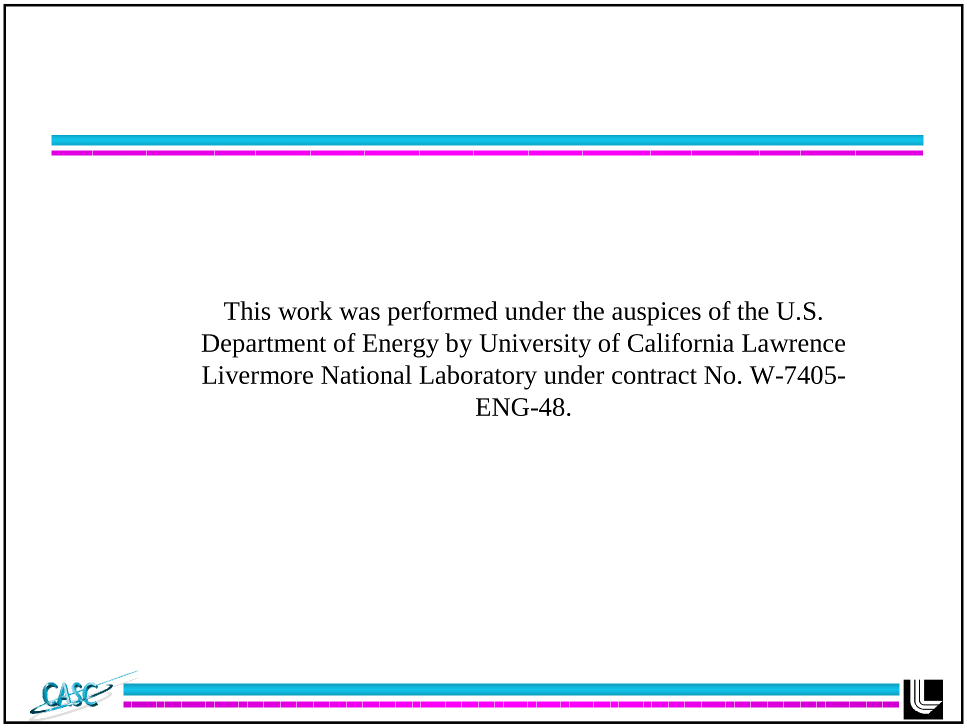
Conclusion

By using XML to define robust specifications of interfaces, we can provide better access to scientific data distributed over the web.



Questions?

www.llnl.gov/CASC/people/critchlow
www-casc.llnl.gov/xewa



This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.